Ovum
TMT intelligence | informa

Ovum Decision Matrix: Selecting a Self-Service Data Prep Solution, 2018–19

Summary

Catalyst

The rise of the self-service analytics era has urged a more flexible, iterative approach to data preparation. The number of data consumers in the typical enterprise has grown massively over time, and over the last decade, business intelligence and visualization tools have increasingly added in guided functionality aimed at further expanding the scope of the user base to nontechnical business users. A higher demand for data sets that are ready for manipulation has challenged IT-centric models for preparing and standardizing data for analysis.

Before widespread adoption of self-service visualization tools, centralized data warehouses were typically used as the "single source of truth" to feed BI tools. Today, this model proves inflexible as the number of data consumers grows, as IT becomes the bottleneck in the data preparation process, relying on IT-centric ETL (extract, transform, load) tools to populate the warehouses. Analytics end users instead need self-service tools for blending and prepping their own data sets, which they can then feed directly into visualization and BI environments; however, IT still needs visibility into the process so that governance can be maintained. Self-service data preparation functionality has developed to respond to this need, providing self-service users with the fluid data access they desire while still permitting IT oversight into the process.

Ovum view

The self-service analytics era demands ready access to prepared data that can be input into self-service visualization and BI tools. The traditional BI approach to data preparation, using ETL tools to populate centralized data warehouses, was never meant to support the modern volume of analytics end users; it was also an IT-controlled process where IT was the bottleneck in preparing data for analysis. As the number of analytics end users has grown, and as users expect ad hoc access to data, users need to be given tools to prep and blend data themselves. Self-service data prep functionality has evolved to give general business users the guided tools they need to transform and prepare data sets without complete reliance on IT, allowing the data prep process to scale within the organization. But as more users access growing volumes of diverse data to prepare for analysis, there needs to be governance of the process. To address this need, many data prep environments are building in native data catalog functionality or providing integrations with best-in-breed standalone data catalog tools so that users can navigate and find the data they need. "Inheritance" of role-based access controls from data source repositories is also important, so that users view only the data they have the permissions to access.

In the market at large, there is a trend of self-service data prep "platformization," whereby data prep functionality is increasingly being baked into larger data management and analytics platforms rather than being offered as a standalone tool. In many cases today, self-service data prep is treated as a feature rather than a product. There are potential benefits to both the platform approach and the standalone, or independent, self-service data prep product approach. A larger platform, with a single stack for more functionality, provides users with a single interface for multiple needs and can make governance of the ecosystem more streamlined. However, the larger the platform, the greater the risk of vendor lock-in; competing interests in terms of functionality can also limit integrations and compatibility with other products. The independent self-service data prep approach typically offers a focus on deep data prep functionality and maximum compatibility with other tools and platforms,

giving the enterprise flexibility. However, the standalone approach's narrow focus on data prep capabilities alone means that other products are often required for core governance and analytics functions. Multiple approaches, both platform and standalone, are represented in this report. Regardless of whether the enterprise selects the platform or standalone approach for self-service data prep, maximum compatibility with data sources – both on premises and in the cloud – should be a key priority in selecting a product.

Guided functionality in self-service data prep tools, driven largely by machine learning, is further expanding the potential user base to increasingly nontechnical business users. As more users join the data prep ecosystem, more features and functionality need to be designed for operationalizing the data prep process within the organization; more users, additionally, compound the need for governance functionality. To operationalize the data prep process within the organization, data prep users need to be able to collaborate and share data prep recipes in order to avoid reinventing the wheel in their work. Today's data prep tools are increasingly becoming fully collaborative ecosystems, with searchable/sharable queries, data sets, and transforms, as well as capabilities for publishing and sharing workflows, pipelines, and other models. Ovum's research suggests machine learning-guided functionality (such as offering predictive transformations and "smart" recommendations for blending data) is becoming a key differentiator between self-service data prep products, and collaborative capabilities will continue to grow in importance as the enterprise tries to govern its self-service ecosystem.

Key findings

- There is an increasing trend of "platformization" of self-service data prep, whereby data prep functionality is being added as a feature in information management and analytics platforms, rather than being offered as a purely standalone tool.
- Leaders in this report represent a variety of approaches: both standalone data prep, as well as data prep functionality embedded within a broader platform. The right choice will depend on business requirements and existing IT infrastructure.
- Self-service data prep vendors are quite closely matched in terms of core data blending and preparation capabilities; differentiators include machine learning-guided functionality, connectivity to analytics tools, and governance features.
- Information governance functionality is increasing in importance for self-service data prep environments, with many products embedding data catalogs; the enterprise demands the ability to audit and see lineage for transformations.
- Because some self-service data prep products are natively embedded in analytics platforms, there is notable variation between products' ability to connect to multiple competing BI and visualization tools.
- Self-service data prep has traditionally served as a feeder to the self-service analytics ecosystem. However, forward-leaning organizations are increasingly feeding prepped data into machine learning models. The only follower in this report supports this unique use case, embedding data prep functionality in a data science platform.
- Enterprise cloud-first and multi-cloud strategies are putting pressure on data prep vendors to offer connectivity to various cloud data repositories and software-as-a-service data sources; most vendors in this report can deploy on all three major cloud providers.

- Native integration with execution environments such as MapReduce, Spark, and Hive give the enterprise flexibility in data processing; a data prep tool's ability to infer best-fit processing execution gives the enterprise an edge in working with large data sets.

Vendor solution selection

Inclusion criteria

Self-service data prep functionality has evolved from a variety of product backgrounds. While standalone self-service data prep tools still do exist, the overarching market trend is to package data prep functionality in data management or analytics platforms. Hence, this assessment needed to look at a wide variety of offerings, since narrowing the market down to purpose-built tools would exclude deep data prep functionality embedded within other environments. At the outset of this research, an inclusion criterion of a standalone or separately licensable offering was listed; however, after initial consultation with vendors, this requirement was removed.

- The vendor must be a global vendor and have customers in at least two of three regions: APAC, EMEA, and North America.
- The focus of the product should be self-service for the business user. The interface should be graphical, rather than code based.
- The data prep product must be able to handle structured, semi-structured, and unstructured forms of data.
- The data prep product must have cataloging and/or metadata repository capabilities available.

Exclusion criteria

While self-service data prep functionality can be found in a number of platforms and environments, the focus needs to be on providing the business user with tools for blending, joining, and transforming data; coding should not be required. This is what differentiates self-service data prep products from more traditional ETL and integration tools. Ease of use is a major focus for self-service data prep vendors, as they compete to add in machine learning-based functionality to guide users in their journey. Exclusion criteria are as follows:

- The vendor's data prep solution requires coding to handle or manipulate data.
- The vendor's solution is more than 50% made up from partner solutions or third-party solutions.
- The vendor has no direct contact with end customers; everything is done through the channel partners.

Methodology

Technology assessment

In this assessment dimension, Ovum analysts develop a series of features and functionality that would provide differentiation between the leading solutions in the marketplace. The criteria groups identified for self-service data prep are as follows:

- **Integration and exploration features:** This refers to the product's ability to connect to various data sources and allow users to conduct initial exploration and profiling of data.
- **Data manipulation features:** Core data prep functionality, including transformation, cleansing, blending, enriching, and modeling capabilities for data, is the focus here.
- **Data governance features:** These are the product's capabilities related to the management and security of data, including metadata management and catalog functionality.
- **User experience and UI features:** This refers to the product's ease of use and range of tools designed for user interaction with the data, product, and company.
- **Administration features:** These are qualities related to deployment, processing, and architecture. Cloud deployment is discussed in this category, as is data processing.
- **Data output and analytics features:** Data prep is a precursor to analytics; features discussed here relate to the export of data, connectivity to analytics environments, and native analysis capabilities.
- **Collaboration and machine learning features:** The focus here is on machine learning-powered functionality and tools provided for user collaboration and sharing of work products.

Execution

In this dimension, Ovum analysts review the capability of the solution around the following key areas:

- **Maturity:** The stage that the product/service is at in the maturity lifecycle is assessed here, relating to the maturity of the overall technology/service area.
- **Interoperability:** In this element, we assess how easily the solution/service can be integrated into the organization's operations, relative to the demand for integration for the project.
- **Innovation:** Innovation can be a key differentiator in the value that an enterprise achieves from a software or services implementation, and this is assessed in this criterion.
- **Deployment:** Referring to a combination of assessed criteria and points of information, Ovum analysts provide details on various deployment issues, including time, industries, services, and support.
- **Scalability:** Points of information are provided to show the scalability of the solution across different scenarios.
- **Enterprise fit:** The alignment of the solution is assessed in this dimension, and the potential ROI period is identified.

Market impact

The global market impact of a solution is assessed in this dimension. In this Ovum Decision Matrix, market impact was measured on revenues alone.

- **Revenues:** Each solution's global self-service data prep revenues are assigned to a revenue band. The highest possible revenue band generates a score of 10, with each lower band generating a proportionately lower score. These scores are used to

size the vendor bubbles on the Ovum Decision Matrix: Self-Service Data Prep charts in Figures 1 and 2.

Ovum ratings

- **Market leader:** This category represents the leading solutions that we believe are worthy of a place on most technology selection short lists. The vendor has established a commanding market position with a product that is widely accepted as best of breed.
- **Market challenger:** The solutions in this category have a good market positioning and are selling and marketing the product well. The products offer competitive functionality and good price-performance proposition, and should be considered as part of the technology selection.
- **Market follower:** Solutions in this category are typically aimed at meeting the requirements of a particular kind of customer. As a Tier-1 offering, they should be explored as part of the technology selection.

Ovum Interactive Decision Matrix

To access the Self-Service Data Prep Ovum Interactive Decision Matrix, an online interactive tool providing you with the technology features that Ovum believes are crucial differentiators for leading solutions in this area, please see the Ovum Interactive Decision Matrix tool on the Ovum Knowledge Center.

Market and solution analysis

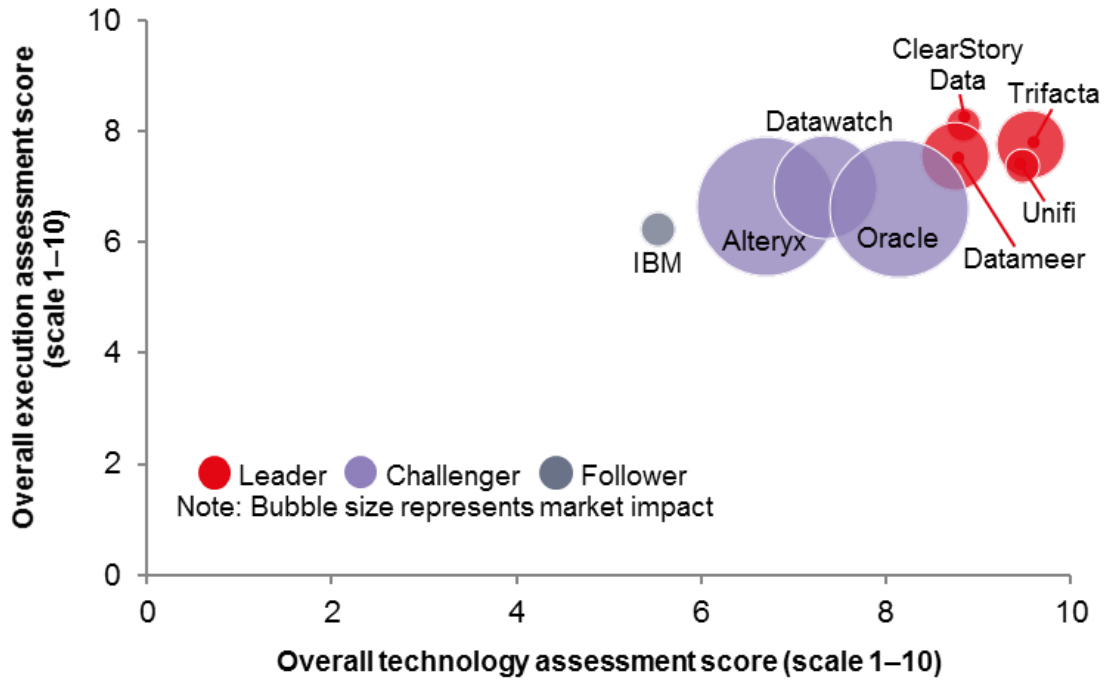
Ovum Decision Matrix: Self-Service Data Prep, 2018–19

The self-service data prep market, paradoxically, is both mature in some regards and still rapidly evolving in others. Across Ovum's technology assessment categories of product functionality, products tended to be very closely grouped in capabilities associated with core data prep functions described by the data manipulation category, such as joins, transformations, and merging, cutting, and replacing of values. Similarly, most products scored highly in the administration category, which encompasses features related to deployment, processing, and architecture; the overwhelming majority of self-service data prep providers, for instance, already support deployment on all three major public cloud providers. The battleground, then, for this Ovum Decision Matrix scoring process came down to categories with more variance in response. These categories, such as collaboration and machine learning and data governance, represent areas of rapid development in the enterprise software market at large; they are not capabilities unique to self-service data prep products, but rather broader trends that are becoming incorporated by necessity within these environments. These are becoming emergent areas of differentiation between products, and vendors are drawing upon their diverse product backgrounds to flesh out functionality in unique ways.

While data prep can be thought of as a discrete step in the data analytics workflow, data prep functionality is increasingly difficult to find as a dedicated, independent software tool. Platforms are generally becoming larger and more multifunctional; however, leaders in this assessment represent a wide variety of approaches and backgrounds, ranging from independent tools to broad data

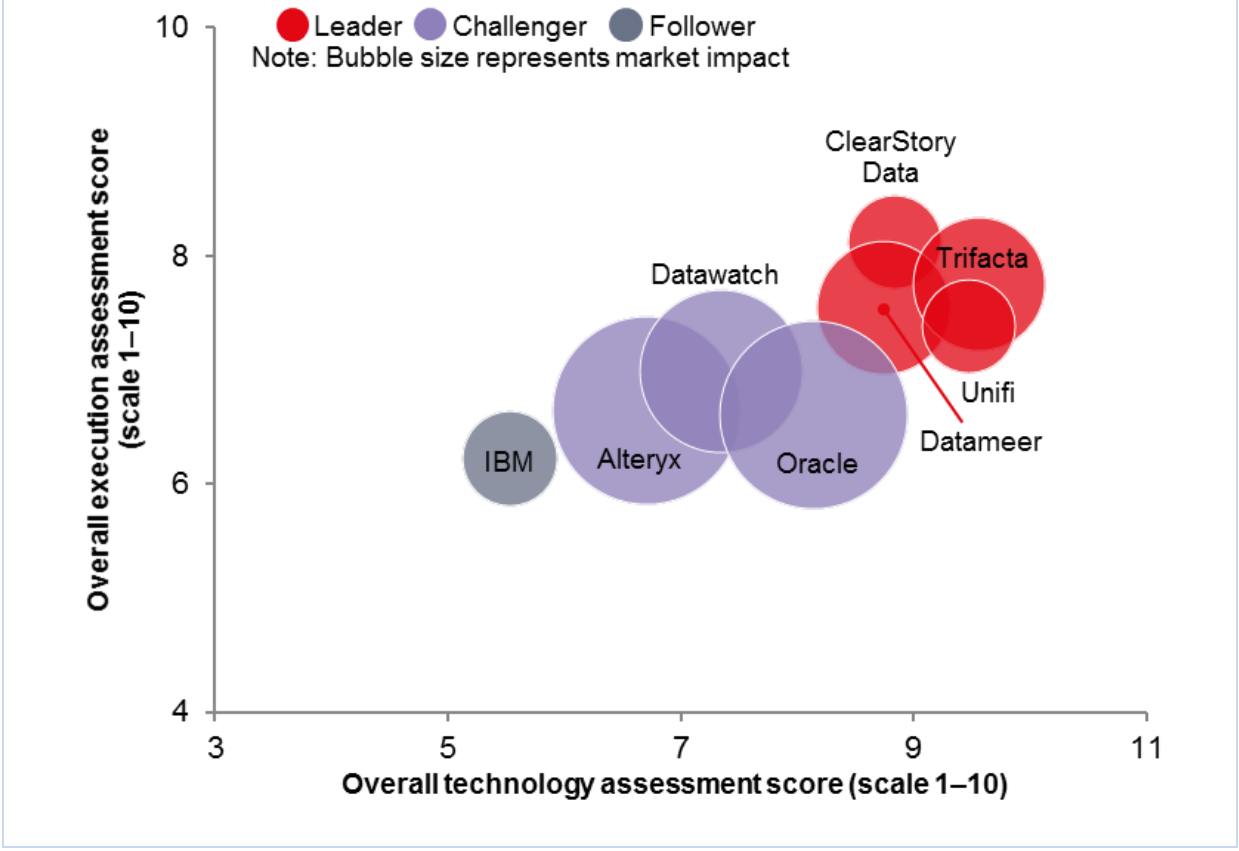
management platforms. This goes to show there is no absolute "best" approach; every organization will have different objectives, business requirements, and existing infrastructure to consider.

Figure 1: Ovum Decision Matrix: Self-Service Data Prep, 2018–19



Source: Ovum

Figure 2: Expanded view of Ovum Decision Matrix: Self-Service Data Prep, 2018–19



Source: Ovum

Table 1: Ovum Decision Matrix: Self-Service Data Prep, 2018–19

Market leaders	Market challengers	Market followers
ClearStory Data	Alteryx	IBM
Datameer	Datawatch	
Trifacta	Oracle	
Unifi		

Source: Ovum

Market leaders: ClearStory Data, Datameer, Trifacta, and Unifi

Market leaders in this Ovum Decision Matrix are ClearStory Data, Datameer, Trifacta, and Unifi. The leading vendors, regardless of their architectural approach, are typically notable for their high scores in the data governance and the collaboration and machine learning categories in the technology features assessment. Additionally, they edged out others in technology categories such as data manipulation, which had closely clustered scoring. For execution categories, market leaders had solutions that tended to score better on maturity and deployment.

ClearStory Data

ClearStory Data provides an end-to-end platform to allow business users to engage in automated data inference, automated data prep, and automated data harmonization. With its intuitive visual user

interface and smart, machine learning-driven recommendations for joining data, general business users (not just analysts) can get up and running with little training or guidance. It additionally includes native visualization capabilities, via its StoryBoards interface, allowing users to explore and discover insights by interacting directly with data.

Datameer

Datameer is a broad and complex platform, originally built on implementations of Hadoop but now native in the cloud, that is focused on building and managing the data pipelines that enable data to be fed into any analytic tool. The current platform performs ingestion, integration, prep, enrichment, exploration, and some visualization. It is well-suited to IT ecosystems that are complex and high-scale. With connectors to more than 70 data sources beyond the Hadoop ecosystem, it provides immense flexibility and connectivity, allowing it to be a central hub for data prep and staging before data is sent to an analytics tool.

Trifacta

Trifacta offers a standalone approach to data prep, fixated on best-in-breed functionality. The company's focus is purpose-built data prep independent of, and interoperable with, a broad variety of BI, data science, governance, and storage and processing environments, both on premises and in the cloud. Its approach is based on a robust strategy of integrations and partnerships. Trifacta's "wrangle once, use anywhere" approach, coupled with its extensive partnerships, integrations, and cloud compatibility, allows the enterprise maximum flexibility and choice in its IT deployments and broad support for a variety of end-user data consumption models.

Unifi

Unifi takes a single-platform approach to address all self-service needs – including data prep – leading up to the visualization and analysis process, providing a single collaborative environment that is bound by consistent application of governance and policies. They offer in one platform an integrated set of capabilities that span four core "pillars" of self-service functionality: governance and security, catalog and discovery, data preparation, and workflow and scheduling. Data cataloging and artificial intelligence capabilities, in particular, are strongly integrated and natively developed as part of the platform.

Market challengers: Alteryx, Datawatch, and Oracle

Market challengers in this Ovum Decision Matrix are Alteryx, Datawatch, and Oracle. They are grouped not far behind the market leaders; as all are mature companies, their "challengers" categorization is overall typically defined more by their lower technology features scores than their execution scores. Their average scores do not tell the entire tale; in several cases, these vendors were brought down in the rankings by a few categories in which they scored disproportionately low on rather than consistently lower scores.

Alteryx

Alteryx offers Alteryx Designer, which provides a bench of more than 250 data tools for preparing and analyzing data from over 80 sources. It has the capability to integrate with additional Alteryx products that provide capabilities around data management (discovery and cataloging), collaboration, and analytics model deployment and model management. Dual options for product use – code-free and code-friendly – ensure that all self-service enterprise users can prep and analyze data equally in the same environment. While Alteryx Designer has an immense following as a data prep tool for

visualization and dashboarding solutions, and it scores highly for product maturity, it is hampered by several of its technology scores such as the data governance and collaboration and machine learning categories.

Datawatch

Datawatch, with its Monarch product, provides purpose-built data prep. It excels at handling semi-structured and unstructured data types, easily extracting alpha, numerical, and date data from documents such as PDFs and incorporating them into the data prep workflow. Usability and breadth of core data prep functionality are two key selling points. By catering to nontechnical business users and power users alike, the product reaches an extended enterprise audience. It scores above average on the data manipulation and collaboration and machine learning technology categories, but suffers disproportionately in the areas of integration and exploration and administration.

Oracle

Oracle is unique in this Ovum Decision Matrix as the only vendor evaluated that tightly embeds its data prep functionality in a dedicated BI, analytics, and visualization platform: Oracle Analytics Cloud. Embedding data prep in the visualization environment makes it easy for users to seamlessly transition from prep to analysis, lending Oracle above-average scores in several of the technology categories. However, this architecture does not play well with other analytics environments, leading to a disproportionately low score on the data output and analytics technology assessment category, which rewards an agnostic approach. As Oracle's data prep capabilities are relatively young, they tended to take the biggest hit in Ovum's execution assessment.

Market followers: IBM

As the self-service data prep market is largely clustered together by scores on data prep technology and execution measures, the determination of follower was made largely by separation from this primary pack, defined by significantly lower technology and execution scores. Market followers tend to be either recent entrants to the market or those that are aimed at meeting the needs of a particular kind of customer.

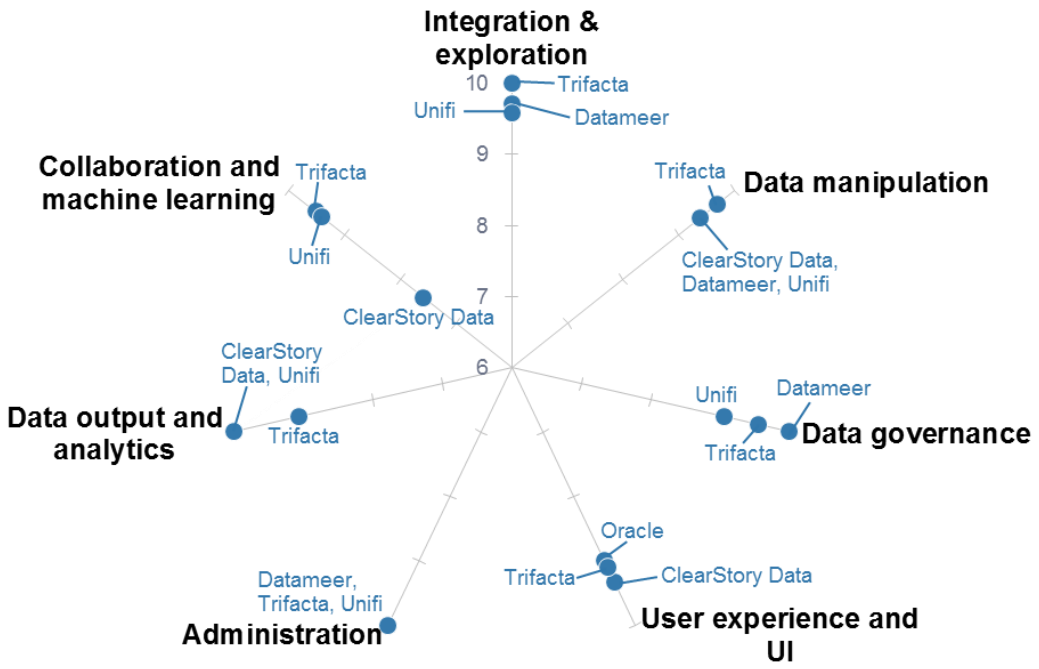
IBM

IBM, the only follower in this Ovum Decision Matrix, is certainly the first and partially the latter. With data prep capabilities that just became generally available in March 2018, it is by far the youngest player in the field and is still rapidly fleshing out core data prep capabilities. It also caters to a unique enterprise audience. As the only vendor in this assessment that embeds its self-service data prep module as part of a data science and machine learning platform – IBM Watson Studio – it is uniquely reaching an audience of forward-leaning organizations that are specifically trying to operationalize machine learning. While most data prep vendors in this assessment are designed with the ultimate end goal of feeding prepped data into self-service visualization and analysis tools, the IBM self-service data prep capabilities are designed to feed directly into machine learning and deep learning models. This gives the IBM product a unique, and noteworthy, approach to consider as a short-list contender for any organization that is looking to scale up data science efforts.

Market leaders

Market leaders: technology

Figure 3: Ovum Decision Matrix: Self-Service Data Prep, 2018–19 – Market leaders – technology



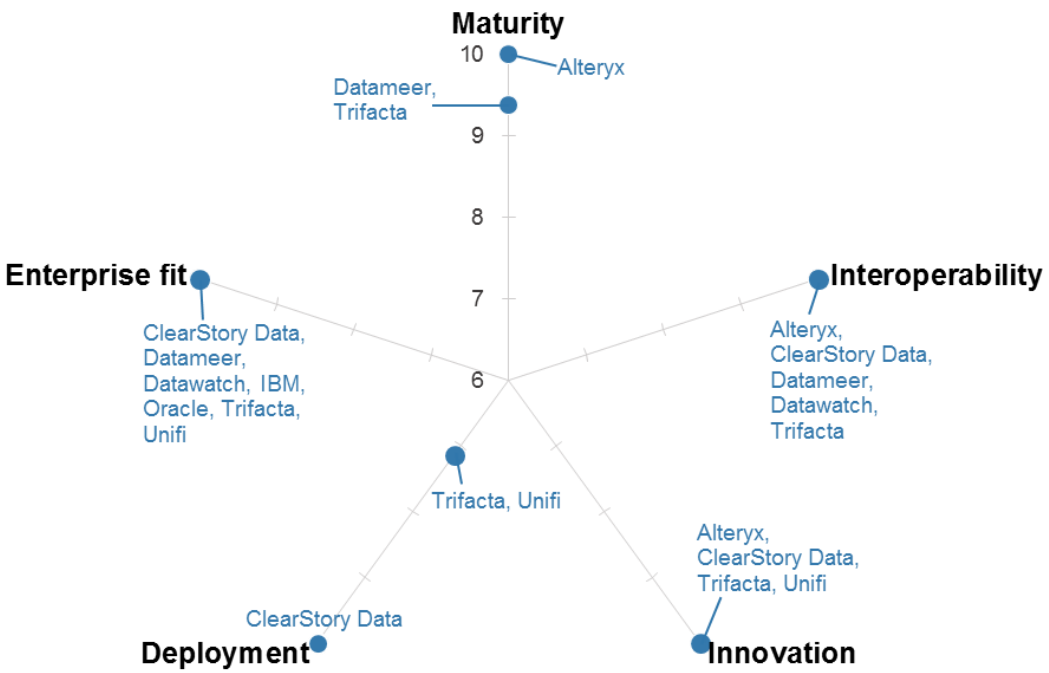
Source: Ovum

Figure 3 shows the top three vendor scores per category in the technology dimension; additional vendors may be listed in the case of a tie for third place. Trifacta, a provider of purpose-built, independent self-service data prep software, dominates these rankings with a top-three score in all seven technology categories. Unifi, a provider that takes an entirely divergent approach with a broad governance and self-service enablement platform, also ranks very highly with a top-three score in six of the seven categories. Rounding out the top three contenders is a tie: Datameer and ClearStory Data, which each score top three in four out of seven categories. Not coincidentally, these four vendors also make up Ovum's group of overall market leaders.

If this chart were expanded to include the fourth- and fifth-ranking scores for each of the technology categories, there would be a lot of close grouping and near overlap. As self-service data prep is relatively mature in many of the technology categories Ovum evaluated, many of the vendors' final category scores were separated by only a few tenths of a point. This was largely the case with the integration and exploration, data manipulation, and administration categories, suggesting relative market maturity in these technical capabilities. Other categories, such as collaboration and machine learning, had a clear group of leaders, followed by a long tail of followers with much lower scores, suggesting these categories are more innovative and indicative of product differentiation.

Market leaders: execution

Figure 4: Ovum Decision Matrix: Self-Service Data Prep, 2018–19 – Market leaders – execution



Source: Ovum

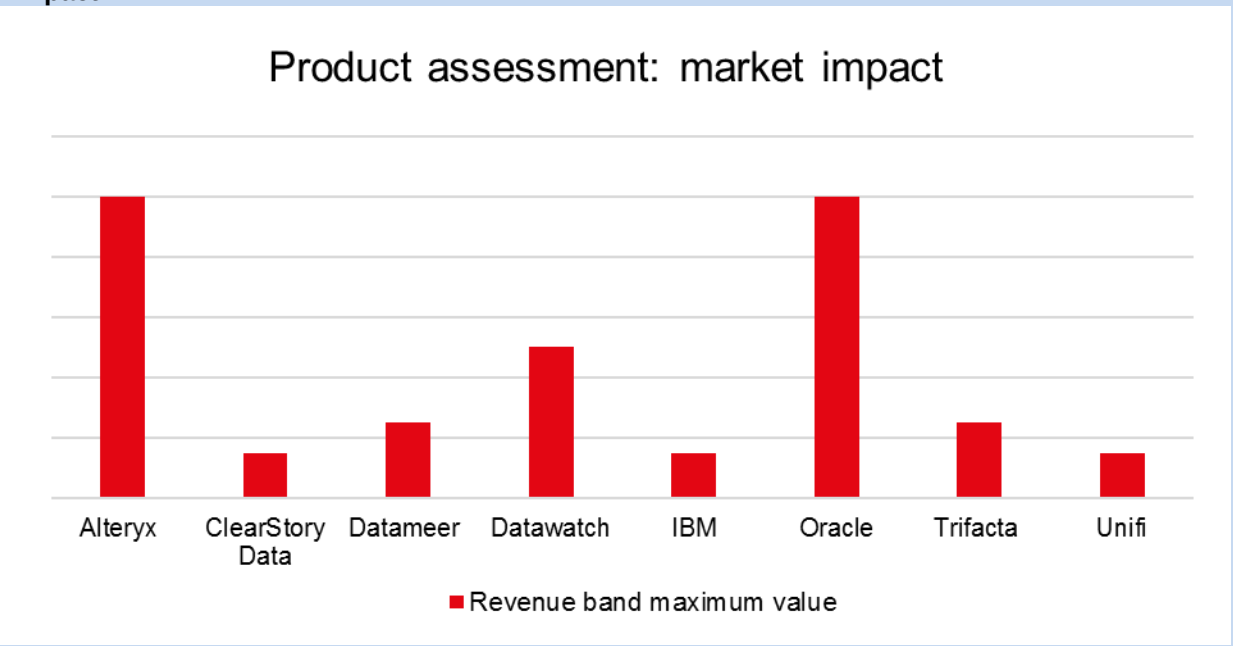
Figure 4 shows the top three vendor scores per category in the execution dimension; additional vendors may be listed in the case of a tie for third place. Trifacta, again, is the dominant player in this dimension, with top rankings in all five of the scored execution categories (a "scalability" category, which was also assessed, was used as supplemental information and did not directly generate a score). ClearStory Data, a provider of both data prep and analysis capabilities, followed closely behind; the company took top 10/10 scores in four of the five execution categories. Alteryx, Datameer, and Unifi all took top rankings in three of the five categories. With the exception of Alteryx, it is these companies that again make up the leaders pack in this year's Ovum Decision Matrix.

As was the case with the technology categories, there was variation in response to the execution categories, with some categories garnering much higher average scores than others. The interoperability category saw five out of the eight vendors scoring a perfect 10, suggesting that this capability (APIs, out-of-the-box integration, etc.) is a mature function of self-service data prep products and is no longer an area of major differentiation. Likewise, the innovation category had four vendors scoring a perfect 10, with several other vendors grouped closely behind, indicating that companies are actively investing in new features and functionality. Enterprise fit saw the least variation, in part due to the low number of scored questions in this category; possible scores were 0, 5, and 10, with the overwhelming majority of vendors scoring the full 10 value. Categories that saw the most variance in response were maturity and deployment. With this report assessing products that range from long-established to brand new, this comes as little surprise. Deployment capabilities –

ranging from SLAs to migration from similar products to industry-specific templates to support options – varied the most between vendors.

Market leaders: market impact

Figure 5: Ovum Decision Matrix: Self-Service Data Prep, 2018–19 – Market leaders – market impact



Source: Ovum

Market impact in this Ovum Decision Matrix was calculated on revenues alone, as many vendors declined to provide all the detailed information required to assess associated measures of market impact such as revenue growth, geographical penetration, and vertical penetration. Vendors were each placed in a revenue band that corresponded to relative revenue derived from their data prep functionality. For the bubble charts in Figure 1 and Figure 2, market impact scores (thus relative bubble size) were calculated based on six possible revenue bands, with a higher revenue band corresponding to a larger bubble size.

In Ovum's findings for this assessment, a higher revenue from data prep – a larger market impact – was not associated with top-ranking scores on the technology and execution measures. The three highest-scoring vendors on the market impact measure, coincidentally, compose the field of the three market challengers in this report. There are several factors that could potentially explain this phenomenon. These three – Alteryx, Datawatch, and Oracle – are simply well-established companies that predate the relatively young self-service data prep movement. Alteryx was founded in 1997, Datawatch in 1985, and Oracle in 1977. All three, logically, started out with offerings that would not be defined today as self-service data prep. But because they had ample time to grow their customer and user base prior to the advent of self-service data prep, they had an existing audience to sell to once these capabilities were developed. This incumbent status and existing user base helped ultimately gain a larger revenue from data prep, despite some of the newer market entrants having more advanced technical functionality and (in some cases) higher execution capabilities.

The newest entrants to the self-service data prep market have had the benefit of learning from these market challengers, which, with the exception of Oracle, began building out self-service data prep

functionality early in the self-service era. Market challengers in this report, in several cases, were the original self-service data prep innovators, whose concepts were built upon by disruptive younger entrants to the market.

Vendor analysis

Alteryx Designer, version 11.8 (Ovum recommendation: Challenger)

Figure 6: Alteryx Designer radar diagrams

Source: Ovum

Ovum SWOT assessment

Strengths

Alteryx is built to equally empower technical and nontechnical users

In most data prep products on the market, power users are given the option of a command line interface. However, this is often the extent of functionality offered to technical users. Alteryx, in contrast, offers a full-featured ecosystem that provides dual options for use. In addition to the product's familiar drag-and-drop workspaces that are built for business users, Alteryx provides tools for running Spark code directly on the Spark cluster (Designer Code Tool for Apache Spark), coding tools for Python and R, and algorithmic support via APIs. A command line option is available for running and scheduling jobs and can be used by people working with the Alteryx scheduler. An API is also provided for monitoring the status of jobs so that admins can script much of the functionality that users normally use the GUI for.

Designer excels at the blending, manipulation, and visualization of spatial data

With Alteryx Designer, it is simple to add spatial context to information. Geocoding and standardization of addresses can be performed, data can be blended on spatial aspects, trade areas can be created and analyzed, drive time can be calculated between spatial data points, and mapping and geographic visualization are available within the product interface. Alteryx's partnerships with leading data providers further augment its spatial capabilities, by allowing data to be enriched by third-party sources. DigitalGlobe provides high-resolution satellite images with street overlays, while TomTom provides visual presentation of over 50 layers of cartographic data.

Weaknesses

Automation and single-click functionality could be more pervasive

Alteryx categorizes the Designer offering as an analytics product. However, it does have well-defined data prep functionality. Compared to some of the data prep capabilities available in other products on the market, Alteryx Designer offers relatively less automation and single-click functionality for core data prep functionality. For novice users, this can steepen the learning curve. Users of Alteryx Designer will find that it does not have automated data deduplication, does not have automated joins, does not have interactive standardization of values, and does not facilitate the iteration over anomalies detected in the generated data set. Data shaping recommendations based on data structure are not offered. There are no automated recommendations for data relationships and keys for combining data across multiple data sets and sources. While data can be enriched by third-party sources such as TomTom, Alteryx Designer does not automatically make recommendations for these third-party sources to enrich particular data sets; the user needs to intuitively know that a data set would benefit from enrichment and seek out the third-party source as appropriate.

Enterprise-grade features and governance require multiple Alteryx products

Alteryx provides two options for deployment between Alteryx Designer and Alteryx Server: the Alteryx Engine can run on the desktop (Alteryx Designer) or be scaled across the organization via Alteryx Server. Alteryx Server provides scaling and governance across multiple teams, including scheduling, workflow sharing, and version control. There is no Alteryx Designer "Enterprise" edition; to get the capabilities associated with enterprise governance and scalability, organizations must purchase multiple products from the Alteryx suite – with the exception of desktop scheduling, which is an add-on to Designer. While all these products are tightly integrated, they do represent separate purchases that some other single-platform data prep products do not require to obtain similar functionality. Data

governance, in particular, depends on the implementation of Alteryx Server, which provides a scalable server architecture that can be deployed on-premises or in the cloud; provides tools for centralized IT administration, monitoring, and reporting; and ensures analytic governance by restricting end-user access to the right data using corporate authentication, permission, and encryption protocols. Data discovery, data cataloging, and collaborative capabilities are found in the Alteryx Connect product.

Opportunities

Visualization enhancements will accelerate data profiling and exploration

While Alteryx has no intention of competing head-on with visualization vendors such as Tableau or Qlik, it maintains that basic visualization capabilities should be a core component of the data prep, blending, and modeling environment. The Alteryx Designer product offers a number of embedded, inline visualization capabilities – "visualytics" – that allow users to examine data more closely before committing to data prep changes or modeling. As Alteryx embeds more visualization capabilities into its data prep and analytic modeling environment, many features and functionality on the product roadmap promise to enhance the user experience and make the visualization features more closely tied to data prep and modeling execution. Currently, visualizations are used primarily as a reference for examining data before committing to changes; Alteryx provides interactive charts for data profiling and robust predictive model support. For the time being, though, interactive charting is available only in the "lab" portion of the product. Still, the company has experience in creating a tremendous number of visual reports, and in the future, the company plans to make all charts fully interactive, allowing users to select elements and directly prompt actions such as data transformations via the chart interface.

Predictive analytics cater to a growing audience of citizen data scientists

Alteryx believes that today's data prep and blending user is tomorrow's citizen data scientist. With data science being a siloed process in many organizations, there is immense value in providing business users with tools that can help scale the data science process and help operationalize it. Alteryx, with Alteryx Designer, has brought in tools built upon the R framework for empowering business users to build predictive models – without code – using the familiar drag-and-drop interface. Alteryx Designer provides 60+ prepackaged tools of the most widely used procedures for predictive analytics, prescriptive analytics, grouping, A/B testing, time series, and forecasting, allowing users to create models and repeatable workflows without programming. Within a single environment, data can be prepped, blended, and used in predictive models to drive business insight, creating continuity in a formerly disjointed process. Alteryx Designer's tight integration with Alteryx Promote proves especially attractive for businesses that are looking to get data science models into production quickly and efficiently without requiring additional development work and DevOps.

Threats

Lack of machine learning-powered data prep functionality may limit user base

When it comes to using machine learning and AI to power existing data prep functionality, Alteryx is somewhat lagging. While the product offers an intuitive drag-and-drop "workflow"-type canvas for joining and manipulating data, there is little in the way of automation or "smart" recommendations to guide users in their data prep journey. Predictive transformations are not offered, there are no predictive search capabilities, there are no automated suggestions or help for actions, and machine learning is not used to identify/formulate repeatable actions and workflows. Overall, there is very little automated or auto-complete guidance for users, and the help that is available (e.g., the "search help"

function within the product) is not powered by machine learning. Machine learning, in general, does not play a role in the assistance of user actions. As the self-service movement extends to increasingly nontechnical users, this lack of automated guidance in the data prep process may limit Alteryx's ability to gain additional, nonanalyst business users.

Governance strategy and messaging could use work to keep pace with the market

Despite being able to connect to more than 80 structured and unstructured data sources, Alteryx Designer, in isolation, does not offer a full suite of enterprise-grade governance tools; Alteryx Connect is needed for functionality such as data lineage, metadata repository, and data catalog. In general, the enterprise must implement Alteryx Server or Alteryx Connect if it expects to obtain a more centralized approach to management and administration of user groups, roles, and settings. Despite this, Alteryx Designer does take a unique approach to data processing that minimizes data exposure. It does not write data to a system to process it, rather writing data to a temporary file on disk. Alteryx deletes all temporary files once data has finished passing through a workflow. The challenge for Alteryx in the increasingly competitive governance era will be clearly articulating the governance and security benefits that do stem from its architecture and promoting its Connect offering as a crucial piece of the puzzle.

ClearStory Data, version 2.7 (Ovum recommendation: Leader)

Figure 7: ClearStory Data radar diagrams

Source: Ovum

Ovum SWOT Assessment

Strengths

User interface is intuitive and geared toward all skill levels

ClearStory Data's user interface is commonly cited by customers as a key strength of the product, and it is tailored with all business users – not just analysts – in mind. The highly visual interface, underpinned by the concept of data "StoryBoards" and "Stories," displays data in interactive charts and graphs, where users can easily drill down, drill across, exert controls on the data, and iteratively

explore aided by machine intelligence. StoryBoards can be commented on and annotated, allowing for a strongly collaborative aspect to the platform, and clicking on a comment automatically shows what the data looked like when the commenter originally posted. Stories are the workbench environment in which transformations and other data prep work is performed, and they allow point-and-click operations. Automation, too, facilitates the speed at which even inexperienced users can get up and running in blending and exploring data. ClearStory Data's data inference automates the inference of semantics in various data sources, reading values, and automating transformations.

Automation and machine learning drive end-user insights

In the ClearStory Data environment, automation is used as a tool to speed the time to insight for the broadest possible audience of users: technical and nontechnical alike. Automation, with underlying machine learning, is used wherever possible within the interface to help users quickly select the best possible course of action in manipulating and handling data, without reliance on technical skills. Automated data inference, which the company considers one of its core strengths, creates a data model using metadata so that users can easily browse the information. Automated suggestions, which are regularly leveraged by end users, can also easily be overridden to help iteratively train the underlying machine learning model. However, the platform does not offer predictive transformations. ClearStory's machine learning-driven knowledge of the data intelligently "scores" related and overlapping data sets to recommend optimal blending strategies on the data at a granular level.

Weaknesses

ClearStory Data messaging needs to reach high-level decision-makers

ClearStory Data's product offers more than traditional data prep; it is an interactive visualization environment that offers deeper insight than data prep tools that simply focus on data joins and transformations for subsequent export to BI and visualization tool environments. However, because of this unique approach that offers capabilities "above and beyond" strictly standalone data prep offerings, it can struggle in the sales process if it is being compared on price and functionality alone to other standalone data prep offerings. This means that ClearStory Data is occasionally dismissed relatively early in the sales process when lower-level enterprise decision-makers are the primary audience. To succeed, it needs to break through to higher-level audiences where its overall value proposition of facilitating enterprise-wide self-service analytics makes more strategic sense.

Enterprise security and governance could use more emphasis

ClearStory Data's goal is to make data and insights more accessible, more quickly, to the largest possible group of end users. In achieving that fluidity and ease of access, emphasis – particularly marketing messaging – around enterprise security, governance, and compliance functionality has been a secondary focus for the company. While the product offers the expected core features associated with security and governance, such as single sign-on, granular permissions, and auditing, some of the more advanced functionality is lacking. Data masking, for instance, is not available, and the product cannot help flag or detect sensitive content such as personally identifiable information. The enterprise that is looking for a data prep platform that comes from a rich security heritage may lean toward some of the larger vendors that have a firmer footing in deep security functionality.

Opportunities

New strategic partnerships expand ClearStory Data's global reach

The core of ClearStory Data's customer base is currently in the North American market, but 2017 saw the addition of seven strategic partnerships that have the goal of expanding the company's global

reach. This expanding partner network will be instrumental in penetrating foreign markets, as the company's relatively small size and US-based offices have previously caused it to focus primarily on the North American market. Now the company is working with strategic resellers and systems integrators to expand its reach globally. This expansion of the partner network comes at a time when the company is seeing significant momentum, and the expansion could act as a multiplier of that momentum into new markets.

Templates can help get ClearStory Data rooted in specific industries

Multiple templates, designed for specific verticals, are available for the ClearStory Data offering, and more are on the development roadmap. ClearStory Data has one of the broadest offerings of industry templates for self-service data prep, helping firms get up and running quickly with their data prep initiatives. By catering to the needs of individual industries, such as specific KPIs and regulatory requirements, these templates can help ClearStory data get a stronger industry-specific following for a product that is relatively horizontal in its market penetration.

Threats

Reliance on Apache Spark limits enterprise processing flexibility

Apache Spark is great at what it does, which is provide a fast and general cluster-computing framework for the processing of big data. Spark is hot now, for good reason, and ClearStory Data's in-memory processing system that tracks the lineage of data sources throughout the analysis process is based on Spark. However, despite Spark's popularity and performance, it is not the only way to process data. As the open source community evolves quickly, there is always the chance that competing – or more performant – projects may emerge. Ability to leverage the native processing capabilities of existing IT resources is also an option that some businesses may want. For the enterprise that is seeking competitive advantage, having more rather than less flexibility and options regarding deployments is almost always preferred. As the open source community continues to evolve, ClearStory Data will need to remain vigilant to evaluate alternative processing engines that may gain traction in the market and add support for these engines when their popularity becomes evident.

Education of the market remains a major challenge

The BI and analytics market is increasingly adding in data prep capabilities to their environments, and previously standalone data prep vendors are expanding their platforms to include more information management capabilities. "Data prep" is no longer a clean, discrete market; it is overlapping, messy, and deeply intertwined with the information governance market. ClearStory Data's visualization capabilities, in particular, require additional efforts at educating the enterprise. Their visualization capabilities are advanced for a platform that is primarily geared toward data prep, but they are not trying to compete with or replace visualization vendors' tools. This can create some confusion with the enterprise audience, which may not immediately see the benefit in paying for overlapping capabilities between their data prep and BI/visualization environments. ClearStory Data needs to provide clear messaging to articulate exactly how its product fits into a holistic BI, visualization, and information management ecosystem within the enterprise.

Datameer Enterprise, version 7.1 (Ovum recommendation: Leader)

Figure 6: Datameer Enterprise radar diagrams

Source: Ovum

Ovum SWOT assessment

Strengths

As a platform, Datameer was originally built for scalability and security

The original functionality of the Datameer platform was self-service data prep on Hadoop. But unlike other self-service data prep tools that initially focused on Excel and other relatively small file inputs, Datameer was preoccupied with the scale and security concerns of enterprise Hadoop environments.

Today, the company's product scales up and out easily. One customer is running the Datameer platform on a 1,400-node cluster. Because it provides a holistic platform for data pipeline management (ingestion, integration, preparation, enrichment, and exploration), this scale is fully warranted. The product integrates directly at the YARN layer of Hadoop and can support data sets into many billions of rows. Furthermore, enterprise security was an early focus of the company. Role-based access controls, support for both NTLM and Kerberos authentication schemes, support for Hadoop secure impersonation, support for SSL, integration with LDAP/Active Directory, integration with enterprise single sign-on, and encryption capabilities round out security functionality.

Datameer excels at data governance capabilities and governance integrations

Datameer is built for purpose on Hadoop (or Amazon EMR), integrating tightly at the YARN level. Partnerships/integration with all three major Hadoop distros (Cloudera, Hortonworks, and MapR), as well as integration with Apache Ranger and Apache Sentry, ensure that all the governance controls that have been implemented in the native Hadoop ecosystem are honored and respected in Datameer's data prep environment, without duplicative administrative work. Data masking and data encryption are available, and the product integrates with LDAP and Active Directory. Row, column, and entity-level security ensure that the right people have the right access to exactly the right data fields. Additionally, the product offers flexible options for data catalog functionality; while there is a native data catalog, there are also integrations available to third-party catalog tools such as Alation. This, in turn, comes with complete audit trail and bidirectional lineage visibility of data prep actions.

Weaknesses

Machine learning and guided functionality are still in relatively early stages of development

Datameer does use machine learning to power several aspects of its functionality: detecting and importing file types, paring semi-structured data, detecting outliers and anomalies, and recommending attributes for data. However, its use of machine learning to provide a guided user experience is still in the relatively early stages. Predictive transformations are offered, but machine learning is not yet used to create automated recommendations for combining data, is not used to identify repeatable actions or workflows, and is not used to help detect sensitive data types. Furthermore, user actions and social activity data are not leveraged to improve the product's machine learning models, as they are in some other products. Given the breadth of the Datameer platform, ensuring a guided user experience will be critical to helping business users navigate increasing volumes of data. Machine learning-guided functionality is the key to this, helping predict the best course of action for users.

Lack of support for streaming data may deter forward-leaning enterprise prospects

Streaming data is a growing part of the enterprise data landscape. There is an enterprise demand for platforms that can simultaneously handle streaming and historical data, and eventually this demand will extend to self-service data prep ecosystems. As Datameer is an analytic data management platform rather than simply a data prep tool, there is certainly value in being able to incorporate streaming data into data pipelines, preparing it for consumption in any tool. Datameer, however, does not have a streaming engine, limiting the data types it can handle. The company self-identifies this as a potential weakness, particularly as forward-leaning prospects and customers increase their focus on streaming use cases. The company is, however, watching the streaming market closely and does plan to add in streaming capabilities when more demand is evident.

Opportunities

Vast data-type compatibility, if maintained, will continue to draw in enterprise customers

Datameer has direct access to a vast trove of data beyond the Hadoop ecosystem, with more than 70 connectors available to various databases and SaaS applications. If a customer has an obscure data source that the platform does not already connect to, Datameer will build out a custom connector via its professional services offerings. As the data lake becomes increasingly virtual, distributed, and hybrid in nature, accessing data in these diversifying locations and sources becomes a competitive advantage; it is in the interest of the enterprise to be able to prep all data in one platform.

Datameer offers this functionality. As long as it can ensure that its development resources are not spread too thinly, and it can maintain its focus on developing new connectors as new repositories emerge, it will continue to attract large enterprise clients that have diverse IT ecosystems. Datameer's challenge with such a broad platform, however, is ensuring that the right amount of development effort is allocated toward integration without sacrificing development of core product functionality.

Cloud flexibility options appeal to the modern enterprise looking to avoid vendor lock-in

In Ovum's assessment of product features, Datameer was one of three companies that scored a perfect 10 for the administration category, which includes cloud deployment options and processing/architectural capabilities.

The Datameer platform can be deployed via public cloud, private cloud, on-premises client, or on-premises server. For public cloud options, the company offers deployment on all three leading providers: Amazon Web Services, Microsoft Azure, and Google Cloud Platform (GCP), allowing the organization to avoid vendor lock-in. As the product supports multiple processing environments as well, existing IT architecture can be leveraged in the most efficient way possible; the Smart Execution decides which engines to use based on the characteristics of the job or parts of the job. This amounts to incredible flexibility for the enterprise in terms of processing, assuming its data architecture is fundamentally based on some form of Hadoop. These capabilities will increasingly be of value to the enterprise, as organizations look to leverage cloud ecosystems and their elastic processing capabilities.

Threats

Datameer is still perceived as being a highly Hadoop-dependent platform

The Datameer platform originally offered a spreadsheet-like data prep interface on top of complex big data sources sitting in Hadoop. While the platform today has expanded to envelope other functions, this core integration with Hadoop and YARN (including Hadoop distros and Amazon EMR) remains central to Datameer's approach. As the data lake becomes more virtual and data storage becomes increasingly distributed, this "dependence" on Hadoop or EMR could be seen as a drawback.

Datameer can access data from Amazon S3 and other cloud object storage repositories; it has connectors to over 70 data sources. But this does not change the fact that the platform essentially needs to run on either Hadoop or EMR. Spark, Tez, or Datameer's own data processing engine can be leveraged under this framework. But for now, the basic dependence on Hadoop and EMR remains (though a product offering for small, single-node data prep environments is available for AWS). In 2018, the company will release a new version with an independent containerized framework using Kubernetes and working with independent object stores, eliminating the Hadoop dependencies. In addition, this same framework will be used for a SaaS offering of the product in the same time frames.

Platform breadth and complexity risks spreading development resources thin

Datameer is a broad and complex platform, focused on building and managing the data pipelines that enable data to be fed into any analytic tool. The current platform performs ingestion, integration,

enrichment, and exploration, and visualization. As Datameer is not a large company, there is a perennial challenge in product development to ensure that resources are allocated appropriately, striking a balance between breadth and depth of functionality – avoiding a platform that's "a mile wide but an inch deep." This occasionally results in trade-offs; for example, Datameer has recently de-emphasized the development of native visualization and infographic features in favor of building out native integrations for third-party BI and analytics tools. The relative complexity of the platform also presents a related challenge: in reaching the broader market, it needs to better communicate its value in more streamlined, simple business use cases, particularly in the cloud.

Datawatch Monarch, version 14.3 (Ovum recommendation: Challenger)

Figure 8: Datawatch Monarch radar diagrams

Source: Ovum

Ovum SWOT Assessment

Strengths

Excels at handling semi-structured and unstructured data types

Datawatch is a well-established company that predates data prep as a market sector; it traces its initial roots to data extraction and conversion, which remains a major component of the data prep process today. Datawatch Monarch was originally designed as a data extraction tool that converted

contents of PDF, .TXT, machine spool data, and others into table format. This heritage forms the basis of the product's dexterity with semi-structured and unstructured data formats today, as Datawatch Monarch can easily detect and take contents, such as alphanumerical and date data, from PDFs and other documents and convert them to structured tables so that they can easily be manipulated, transformed, and joined with other data sources in the data prep process. Additionally, the product has web-scraping capabilities, making it simple to harvest valuable data directly from internet sources, which makes it immediately available for use in data prep without any intermediary steps.

Datawatch Monarch has all basic and advanced data prep functionality

When it comes to core data prep functionality, Datawatch Monarch covers all the bases. The product functions as a purpose-built data prep tool that can perform everything from relatively basic joins (inner, outer, left, and right) all the way to complex user-defined calculations and enrichment of data from third-party data sources. The product makes it easy for users to iterate over any anomalies detected in the generated dataset, and it uses fuzzy matching for detecting approximate matches between values. Additionally, the product conducts a range of functions automatically: deduplicating data, inferring data structure/schema from the data source, and converting source format to destination format. Single-click functions help users easily append and join data sets. Because the product can join and merge data across multiple data sets and sources, it is an ideal option for organizations that need to prep data from various different repositories. Datawatch Monarch also caters both to nontechnical and power users alike; users can navigate from a business-oriented graphical interface, or dive even deeper using a command line interface. There is very little that the Monarch product is lacking when it comes to functionality for exploration, transformation, cleansing, blending, enriching, or modeling of data.

Weaknesses

Native in-memory processing allows for little processing flexibility

Datawatch Monarch employs a native in-memory processing engine for executing the transformations and other manipulations of data that are associated with the data prep process. This native in-memory processing has benefits: processing for data prep functions occurs automatically, and the enterprise has no need to select or configure a processing environment. However, in receiving this streamlined functionality, the enterprise faces a trade-off; Datawatch Monarch allows for little to no flexibility in employing alternate processing engines or environments. As open source processing frameworks such as Spark gain popularity, the enterprise benefits from data prep tools that allow the flexibility to pick and choose processing environments based on scale, data types, and other variables. Datawatch Monarch does not support these alternate processing environments, though eventual support is on the development roadmap.

Hadoop integration could be more extensive, particularly for governance

Datawatch Monarch has direct connectivity and integration with the Hadoop Distributed File System (HDFS), which allows it to access data stored in the Hadoop ecosystem. Furthermore, integration with Cloudera helps maintain governance over data for organizations that use Cloudera as their Hadoop distribution. However, Datawatch Monarch's integration with Hadoop occurs only at the HDFS level and does not leverage some of the available security and governance frameworks that are available for the Hadoop ecosystem. Datawatch Monarch's integration with Hadoop does not occur at the YARN level of the Hadoop stack, and integrations with Apache Ranger and Apache Sentry are not

available. The net effect of these absences is that Datawatch Monarch does not integrate with the security and governance features of the Hadoop ecosystem as well as it possibly could.

Opportunities

Monarch Swarm provides opportunities for land-and-expand growth

Monarch Swarm, which was released in 2017, is the company's add-on data management and social platform for the governance of the self-service ecosystem. What Monarch Swarm aims to do is provide centralized, controlled access to data that is governed by data stewards and enriched by social activity. Users can "like" data sets, follow individual users, curate data sets, comment on data sets, and be alerted when new data sets are created, helping eliminate duplicative effort. This is combined with a centralized data marketplace, which breaks down silos by giving users a single place to "shop" for raw and/or curated data sets, reports, models, workspaces, and expertise. Monarch Swarm has also given Datawatch the opportunity to modify its go-to-market strategy. The development of Monarch Swarm opens the doors for an aggressive land-and-expand strategy for the company; once the desktop version of Monarch has been installed in the enterprise, Datawatch has a foot in the door to introduce Monarch Swarm's deeper social and governance capabilities, which allow for a managed self-service ecosystem.

Machine learning capabilities are strong, and strategy still has room for growth

The Datawatch Monarch product uses machine learning for a litany of purposes. Recommendations for combining data, suggestions for user actions, detection of outliers and missing values, recommendations of attributes to increase the contextual value of data, and identification of repeatable workflows/actions are all powered by machine learning algorithms. The user ecosystem of the Monarch product itself also helps sustain the machine learning paradigm; user actions and metadata from user social activity are all used to help improve machine learning models over time. The more the product is used, the better the machine learning functionality becomes. However, there is still room for growth. Predictive search and predictive transformations, in particular, are areas where machine learning has yet to be applied. The company has not yet publicly put a prominent emphasis on its machine learning strategy, though it certainly has the opportunity to do so.

Threats

Visualization vendors could end up eating a piece of the data prep pie

For all of Datawatch Monarch's strengths in core data prep functionality and dexterity in handling semi-structured/unstructured data, it is not a visualization tool. There are no automated best-fit visualizations for data sets, and there are no interactive visualizations where actions (such as clicks) can prompt transformations or other data prep functions. Visualizations from BI/analytics tools cannot be embedded within the Monarch environment. Instead, Datawatch's visualization strategy is centered on partnerships and integrations, with the company offering integration and native file format compatibility with a wide range of leading BI and analytics tools. However, some data visualization and analytics vendors are beginning to move aggressively into the data prep space by adding in basic functionality such as those for transforms and joins. Datawatch has not yet seen this trend pose a material threat to its business, because Monarch provides functionality beyond acting as a mere "feeder" to analysis tools. But with the self-service analytics trend continuing to grow, these visualization ecosystems will continue to gain more users, creating undeniable gravitational pull for native data prep capabilities.

Data cataloging strategy needs more clarity to compete in a growing market

Datawatch Monarch, in its present state, has its own natively developed data cataloging capabilities that catalog data sources, data lineage, relationships between data, and transformations that have been applied to data. However, the company is also planning potential integration with third-party cataloging tools for 2018. While integrations provide more options for the enterprise, they also potentially create confusion around Monarch's existing capabilities and strengths. As Datawatch opens the doors for integration with third-party catalog vendors, it will need more clarity around the messaging and value proposition of its own native capabilities, which may overlap with those of third parties. What Datawatch will need to do, moving forward, is to play an active role in helping the enterprise understand which catalog option is best for specific business requirements, without underselling its native catalog capabilities, which are tightly embedded within the Monarch environment.

IBM Data Refinery (Ovum recommendation: Follower)

Figure 8: IBM Data Refinery radar diagrams

Source: Ovum

Ovum SWOT Assessment

Strengths

Platform approach unifies data prep, data catalog, and data science

Data Refinery's current packaging within the IBM Watson Studio and IBM Watson Knowledge Catalog ensure that users of either environment have the tools they need to access and prep data as needed without toggling between products. The user experience and UI of IBM Watson Studio and IBM Watson Knowledge Catalog is unified in a single environment, lending a single, seamless experience

and frictionless access to all data manipulation capabilities, including data catalog and exploration, data prep, data science, and even preliminary visualization and analysis. Because data consumers within the enterprise are united under a single platform, with access to all the tools they need, governance is increased and the ability to operationalize the data exploration and data prep process is bolstered. IT has direct visibility into the platform's users and activities, offering a single point of control for important functions such as role-based access controls and data policies. Unifying data catalog, data prep, and data science workbench capabilities into a single environment meets an important need: serving the widest possible array of self-service users in the same cohesive ecosystem, where they can be managed and monitored by IT together.

Governance capabilities bolster compliance and help meet regulatory needs

IBM's Data Refinery capabilities, as packaged within the IBM Watson Studio and IBM Watson Knowledge Catalog, are compliance-ready out of the box. Aside from providing a complete audit trail, the product provides a number of governance and security capabilities that bolster compliance with various data protection regulations, such as the EU's General Data Protection Regulation (GDPR). Granular, role-based access controls, inheritance of role-based access controls from data source repositories, and row/column/entity level security all ensure that data is viewed and accessed only by authorized parties. When users search for data in the product interface, only results that they are authorized to view are returned. Additionally, having a single platform for data catalog, data science, and data prep inherently consolidates governance control and IT visibility; one platform (as opposed to disparate tools) means more streamlined monitoring, policy enforcement, and auditability. But perhaps the most powerful tool in Data Refinery's arsenal for compliance is its machine learning-powered capabilities to automatically detect and mask data that is potentially sensitive, such as Social Security numbers and phone numbers.

Weaknesses

Direct integration with self-service BI/visualization tools is underdeveloped

In all, the capabilities of the IBM Watson Studio are designed for governance, data science, and data prep, all from a single UI and platform. It would follow, then, that self-service analytics and visualization would be the logical conclusion of these platform capabilities, with users completing data prep functions then being able to fluidly transition to full-powered visualization and analysis with functionality for direct integration into the native IBM analytics environment (IBM Watson Analytics) or preferred third-party BI tools. However, this is still not the case. Once data prep has been completed in the IBM Watson Data Platform environment, the process for getting data into self-service BI and analytics tools is still relatively manual in nature. Direct connectors for popular visualization tools, such as Tableau and Qlik, have yet to be developed in the IBM Watson Data Platform environment (although a direct connector to Tableau is planned for early 2Q18), and native support for partner file formats (e.g., .tde, .qvd, .pbi) is not yet offered.

Features for automation and guidance of the user journey are currently lacking

Automation is an area where Data Refinery does not perform as strongly as some of the other data prep players, perhaps because of its relatively recent entry into the market. In general, users need to know what they want to do with the data; predictive transformations are not offered, and automated suggestions and help for actions are not provided. Other automated and single-click functionality is lacking; automated visual flagging of outliers, anomalies, missing, or mismatched content is not conducted, data is not automatically deduplicated, there are no automated joins, and there is no single-click function to append or join multiple data sources. The product does not offer automated

recommendations for data relationships and keys for combining data across multiple data sets and sources. While data can be enriched with third-party sources, sources to enrich data are not automatically recommended: users need to seek them out.

Opportunities

Data prep capabilities are well-positioned to accelerate data science initiatives

By embedding the Data Refinery directly in IBM Watson Studio, IBM has positioned its data prep capabilities not only to help accelerate traditional analysis and visualization efforts by business analysts, but also to help data scientists refine the data that they need to build models and get them to production. The decision to offer data prep capabilities directly in the Watson Studio environment reflects IBM's holistic approach to operationalizing data science within the enterprise. By giving data scientists a centralized workbench of tools where they can access notebooks, manage data science projects, schedule analytic compute runs, and manage access and tracking lineage to different sources of data, IBM is providing a platform that "unsilos" the data science process from individual laptops and machines. With the Data Refinery directly embedded in Watson Studio, data scientists benefit by being able to quickly and easily blend and refine the data they need for input into models, without ever leaving the product interface. Data prep capabilities make Watson Studio more comprehensive, helping facilitate adoption and making the product "stickier" for users.

Single-platform ecosystem can potentially operationalize the analytics process

What the Data Refinery may lack in automation and guidance of the end-user journey, it makes up for in its single-platform approach: IBM Watson Studio, under which the IBM Watson Knowledge Catalog and IBM Data Refinery exist, encompasses a broad range of data-handling functions. IBM Watson Studio is meant to be a one-stop shop for the leverage of data, with functionality built for data scientists, business analysts, and data engineers. Providing a single platform under which all these personas can work side by side, with access to the tools they need, helps operationalize the analytics process by eliminating organizational/product silos and centralizing data governance functions. One platform, with one UI, simplifies use and increases adoption. The missing piece here, which IBM is actively working on, is direct connectivity to BI and analytics tools (discussed in "weaknesses," above).

Threats

Late entry to the self-service data prep market means there is catching up to do

The Data Refinery module in the IBM Watson Studio and Watson Knowledge Catalog is a relatively recent addition to the product ecosystem; private beta was initially made available in October 2017. The capabilities are embedded rather than standalone, giving the company a head start in developing the necessary connectors (particularly to IBM data sources) and tapping into the existing collaborative structure of the Watson Studio to build out tools for publishing and sharing of workflows, pipelines, and other models. IBM was not building a self-service data prep ecosystem from scratch, but rather adding self-service data prep tools to an existing data management and collaborative platform. However, this relatively late start to adding self-service data prep functionality into the environment means that it faces an uphill battle in catching up to other data prep providers, particularly those that offer standalone functionality.

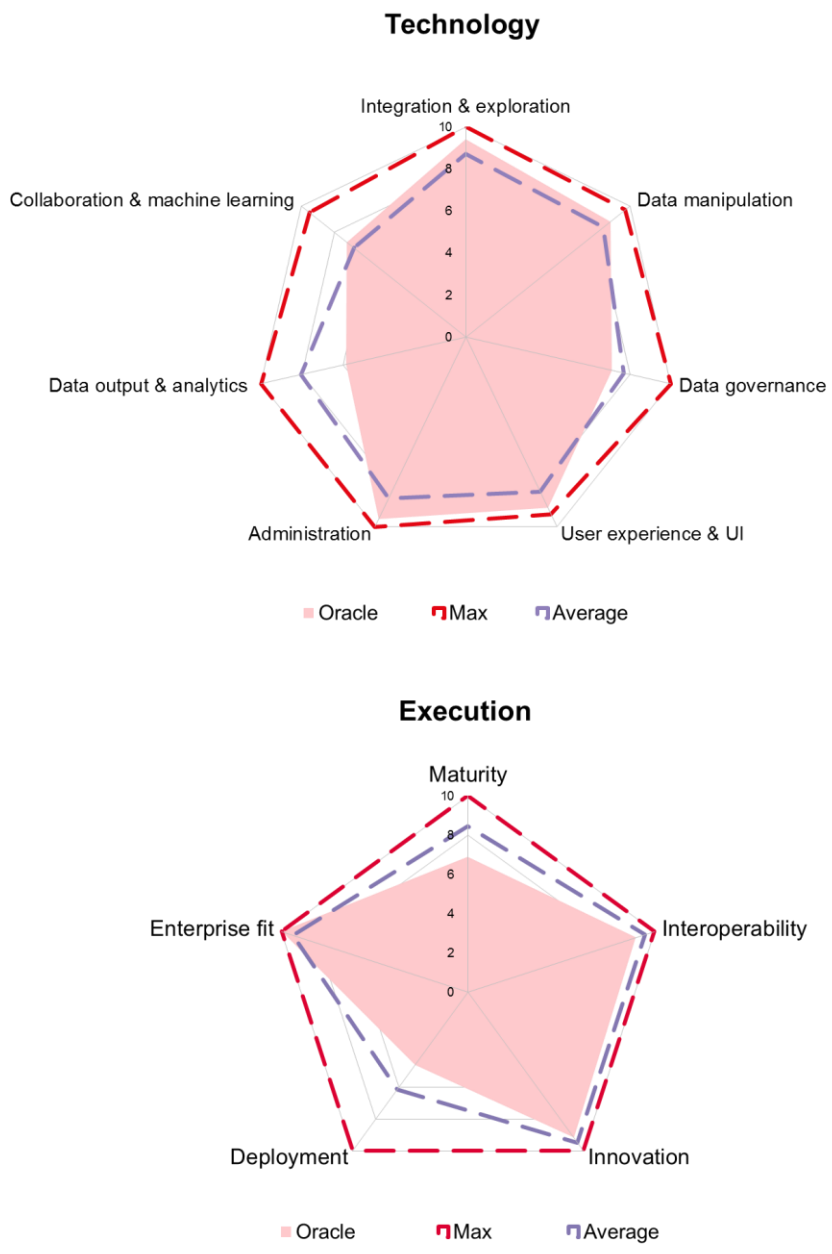
Ecosystem of data connectors needs to expand to keep pace with market

Users of IBM's Data Refinery capabilities have access to 30+ connectors to IBM, non-IBM, and third-party data sources. These connectors should take care of the bulk of data access needs for most

organizations, whether data exists on-premises, in the cloud, or on the desktop. Connectivity to Hadoop (HDFS), Amazon S3, traditional databases (RDBMS), and NoSQL sources cover a wide range of data – particularly big data – that the typical enterprise may have in various locations. However, with the number of repositories and data sources only becoming more diverse within the enterprise, IBM will need to quickly scale up the number of connectors that it offers to keep pace with the market; some vendors of data prep functionality support connectivity to 80 or more unstructured and structured data sources.

Oracle Analytics Cloud, version 17.4.5 (Ovum recommendation: Challenger)

Figure 8: Oracle Analytics Cloud radar diagrams



Source: Ovum

Ovum SWOT Assessment

Strengths

Data prep capabilities are tightly integrated with the Oracle ecosystem of apps

Oracle's data prep capabilities are deeply embedded in the Oracle Analytics Cloud and are available not as an add-on, but rather to all purchasers of the product. This reflects Oracle's philosophy that data prep is not a unique market of its own, but rather something that is inherently tied to the data analytics process; you cannot visualize and analyze data unless it has been adequately prepped first. The flow of data from the data prep environment to the visualization and analytics environment should be seamless and effortless, in the same interface, and not just integrated. Oracle Analytics Cloud delivers this, but also inherently embeds tightly within the entire ecosystem of Oracle apps, making it easy to access and prep data from Oracle favorites such as enterprise resource planning, enterprise performance management, human capital management, and supply chain management. Simple access to data stored in Oracle databases is a given, and Oracle Analytics Cloud automatically inherits all the governance controls – such as role-based access controls – that have been set at the repository level.

Product gives full flexibility in processing and cloud data access options

As a company, Oracle has made the cloud front and center of its business strategy. Oracle Analytics Cloud is deployed only on the Oracle Cloud, but it can access data from any place, or any cloud. This array of options allows customers to leverage data residing within their existing IT investments and infrastructure, without having to lock entirely into the Oracle ecosystem. Because the product can access data from any cloud environment, it is well-suited to any organization that is forging a cloud-first business strategy. Cloud deployments may be public or private, and the enterprise always retains the option to deploy on-premises. Flexibility in processing, too, is a strength of the Oracle Analytics Cloud data prep environment. When prepping data at enterprise scale, organizations benefit from being able to dictate the processing environment based on factors such as data scale and types of transformations being applied. Additionally, the product offers best-fit processing execution; determined based on data scale and types of transformations being performed, the product automatically infers the best-fit processing environment and executes the job based on these factors.

Weaknesses

Those seeking compliance-ready data prep will find features lacking

For organizations that fall under regulatory control, the data prep process must be treated like any other data-handling process; it must be fully transparent and auditable. Oracle Analytics Cloud, however, is lacking in some of the features that would make it compliance-ready out of the box. At this time, a complete audit trail capability is not available, and data lineage (with bidirectional visibility) is available only as an add-on to the standard product package. While the product's native data cataloging capabilities do track data sources, relationships, and transformations, the data catalog alone is not sufficient in fulfilling regulatory requirements for auditability.

Options for integration with non-Oracle visualization tools are limited

Oracle Analytics Cloud is an extraordinarily powerful analyzation and visualization environment for data, with extensive use of machine learning functionality to guide users, including those who are nontechnical, through the process of data exploration and data prep. The data prep capabilities that are embedded in the environment are purpose-built to feed data directly to the analysis features. Users of Oracle Analytics Cloud should not, in theory, need to use another data visualization tool

because the product has all the capabilities one would expect from a leading analytics and visualization platform. However, the flip side of this embedded relationship is that Oracle's data prep capabilities do not play well with other visualization environments; it is purpose-built data prep for the Oracle Analytics Cloud environment, and those organizations with other visualization tools will likely find that the Oracle ecosystem has limited connectivity to these products. Oracle Analytics Cloud does not offer native file format support for other analytics files, such as .tde, .qvd, or .pbi.

Opportunities

Machine learning and automation can prove to be key differentiators

Machine learning in the context of Oracle's data prep capabilities cannot possibly be discussed without mentioning the broader machine learning capabilities of the entire Oracle Analytics Cloud environment. Oracle has made automation and machine learning a core philosophy of development of the Oracle Analytics Cloud, powering automated visualizations and guided data exploration. Machine learning is used to facilitate data prep in myriad ways: detecting and importing files, parsing semi-structured data, creating recommendations for combining data, creating user suggestions and predictive transformations, detecting outliers and anomalies, and detecting sensitive data types such as SSNs. These capabilities blend seamlessly with the machine learning-driven recommendations and guidance offered in the broader Oracle Analytics Cloud environment, creating a single environment for visualization and data prep that is powered by smart recommendations.

As cloud adoption grows, SaaS connectors will be a competitive advantage

Oracle Analytics Cloud, as a full-service analytics product, has an unusually rich ecosystem of connectors to SaaS sources. Connectors to common applications such as Salesforce and Marketo are a given, but also include more obscure sources. Given that Oracle's data prep capabilities are embedded in the Oracle Analytics Cloud environment, this means that data prep users and data prep functionality have equal access to these connectors and sources. Customers frequently cite the ability and ease to connect to SaaS sources as one of the primary selling points of the product. This is especially true in large enterprise deployments where a large ecosystem of SaaS apps is in use.

Threats

Pure visualization vendors are quickly building out data prep functionality

Visualization and analytics vendors are increasingly building out basic data prep functionality, and these embedded capabilities have enterprise appeal because users do not need to toggle between products in order to conduct data prep and analysis. While the same is true for the Oracle Analytics Cloud environment – data prep and analysis are available in the same environment – it is the business model of the pure visualization vendors that makes them a threat to Oracle. The land-and-expand strategy has the potential to infiltrate the enterprise from the ground up, converting knowledge workers to loyal users along the way. Once established in an enterprise environment, these visualization tools often offer data prep capabilities that are temptingly "good enough" for most use cases, creating gravitational pull. One way that Oracle can buffer against these players is by continuing to ensure a very high ease of use, making the product intuitive and sticky for users.

Reaching beyond the existing Oracle customer base may have challenges

Existing Oracle customers make for a natural audience for the Oracle Analytics Cloud. However, to achieve market leadership, Oracle will have to expand this audience to include organizations that do not necessarily have existing Oracle deployments. The challenge facing Oracle is that it is not selling standalone data prep; it is selling a full-service BI and analytics platform, complete with data

cataloging and governance capabilities. This means that the wide swath of the enterprise market that already uses a BI or visualization tool will find no use for Oracle Analytics Cloud, unless they are to rip and replace their existing deployment. As the data prep market is less mature than the market for BI and visualization tools, there is a relatively large market for standalone data prep tools that can complement existing BI and analytics deployments.

Trifacta Wrangler, version 4.2 (Ovum recommendation: Leader)

Figure 8: Trifacta Wrangler radar diagrams

Source: Ovum

Ovum SWOT Assessment

Strengths

Purpose-built approach puts focus on data prep functionality and usability

Trifacta has an extremely focused approach; the company offers purpose-built data prep and best-in-breed product integration for all auxiliary functionality. This philosophy allows the company to pursue two primary objectives: delivering the best possible data prep functionality, with the best possible usability. Usability and best-in-breed data prep functionality make the product sticky for end users, and tight integrations with information management architecture and platforms effectively embed the product into existing governance frameworks. This integration methodology allows Trifacta to focus on core data prep capabilities; Trifacta can typically deliver deeper, richer data prep functionality to its product faster than vendors with broader platforms and a less-focused development strategy. It also allows customers to leverage best-in-breed solutions to flesh out their data governance ecosystems rather than relying on mediocre, natively developed capabilities in tools that do not work well outside their own stack.

Strong governance and integration capabilities maintain control over data

Trifacta's laser focus on core data prep functionality, and the decision to not spread itself too thin in terms of functionality, have clear benefits for the governance of data. Because of its tight integration with platforms and frameworks, Trifacta allows the enterprise to integrate the data prep process into existing information management architecture. Integration with LDAP and Active Directory, integration with single sign-on, support for both NTLM and Kerberos authentication schemes, and support for SSL make Trifacta especially well-suited to secure enterprise deployments. Full data lineage, with bidirectional lineage visibility, ensures an audit trail for compliance and provides the ability to restore data to former states. Trifacta Wrangler is not trying to be everything to everyone in terms of data governance capabilities; instead, it focuses on integration with a wide variety of best-in-breed solutions that perform governance functions.

Weaknesses

Some features are delivered solely through integrations with a data catalog

Trifacta's philosophy is that a neutral data cataloging platform can provide more data type and application compatibility than proprietary data catalog features that are bundled in a data prep platform. Because Trifacta purposely chooses not to develop any of its own catalog capabilities, this places the onus on the enterprise to select and implement a data catalog that is best suited to its needs. Trifacta addresses this by offering an industry-leading array of integrations and partnerships with data catalog tools, giving the enterprise flexibility in choice. However, the enterprise that does not have a third-party data catalog in place when using Trifacta will be missing out on some of the potential functionality of the Wrangler platform. Natural language search, which is a powerful facilitator of usability for nontechnical business workers, requires bidirectional integration with a data catalog and is not available in isolation. Additionally, some of the social functionality of the Wrangler environment, such as support for hashtags (#) and mentions (@), as well as the ability to use machine learning to determine the social news feed, are delivered through integration with a data catalog.

Trifacta needs to accommodate an increasingly global business strategy

Trifacta has an impressive global customer footprint, with users of all versions of Wrangler (including the free version) existing in more than 10,000 companies from 143 countries around the world. The

company is actively looking to expand, particularly in the Asia-Pacific region, where it currently works through resellers and distributors. However, with a global presence and growing global ambitions, Trifacta needs to ensure that both the product and company can meet the needs of these diverse markets. One area in which deficiencies are most evident is Trifacta's lack of options in UI localization; the product is available only in an English interface, despite the wide array of international users and customers. Additionally, Trifacta could use more sales support in the foreign markets – specifically Asia-Pacific – in which it is trying to expand.

Opportunities

Strengths in cloud compatibility poise the company for cloud growth

Trifacta, until recently, was mostly deployed by the enterprise on premises. However, over the past year, there has been an immense swell in customer demand for cloud implementations of the product, and the company's strategy has been extremely forward-leaning with regard to development of cloud compatibility and the ability to seamlessly port "recipes" between different cloud and on-premises environments. Available across all three major cloud providers – Amazon Web Services, Microsoft Azure, and GCP – Trifacta offers customers immense flexibility in how they deploy the product. Trifacta works extensively with the three leading cloud providers to ensure a high level of support and compatibility; rather than just ensuring deployment on the cloud environments, Trifacta seeks out certification and accreditation. Trifacta was recently selected as a Tier-0 partner for Microsoft, the highest tier available, securing it dedicated Microsoft headcount and resources. Amazon awarded Trifacta AWS Machine Learning Competency in late 2017, and the Wrangler Edge and Wrangler Enterprise products are available on the AWS Marketplace. With Google, Trifacta has an OEM relationship, powering the Google Cloud Dataprep solution.

Operationalization of data science can increase demand for efficient data prep

Trifacta offers numerous strategic partnerships and integrations with data science tools and platforms. This gives it an edge in the messy and rapidly expanding data science market, where a multitude of tools, products, and languages means that integration and compatibility is critically important. Trifacta's aptitude in ensuring integration and partnership will prove especially valuable as the data science market grows and diversifies. The company already has tight partnerships with fast-growing upstarts in this space, such as DataRobot and Domino Data Lab. Large enterprise customers have also found Trifacta is complementary to existing software investments in SAS and R. Data scientists want flexibility and choice in the tools they use; therefore, it will be up to data prep providers to ensure that they are compatible with these ecosystems. Trifacta is ahead of the curve in providing integrations and partnerships in the data science space, providing them with a growing opportunity as data science becomes more established within organizations.

Threats

Lack of data catalog may provide opportunity for other vendors to encroach

Data prep vendors with native data catalog functionality argue that embedded capabilities make for more rigorous governance of data, as policies can be consistently applied in one environment. Additionally, some organizations may perceive a one-stop-shop approach to data cataloging and data prep to be easier to manage than two (or more) separate products. While an independent catalog ultimately might offer more extensive functionality and flexibility, the native data cataloging capabilities of many data prep environments might be deemed sufficient for most use cases, creating a gravitational pull for these products. This is especially true for smaller and medium-sized

organizations that may not have yet had experience with an independent data catalog solution. Trifacta needs to pay special attention to potential customers that do not already have a data catalog and help them understand the value proposition of using an independent catalog environment.

Developing integration with new products and platforms is an ongoing race

Because of Trifacta's focused approach to core data prep functionality, its ability to seamlessly integrate with other products and platforms is critical to embedding it in enterprise IT infrastructure. The company offers one of the broadest ecosystems of partnerships and integrations available for a data prep tool, which helps the enterprise deploy the product in a multitude of flexible ways. By focusing on core data prep functionality and not trying to compete with products that offer related capabilities, such as visualization or data cataloging, Trifacta positions itself as a "Switzerland" of data prep, capable of partnering and integrating with the widest possible variety of technology providers. This allows the enterprise ultimate choice and flexibility in architecting its IT ecosystem with a variety of products. This reliance on partnerships and integration, however, means that Trifacta is always in a race to develop new integrations as appropriate based on market developments. Trifacta depends, competitively, on the viability of these integrations more so than competitors that have broader platforms.

Unifi Data Platform, version 2.6 (Ovum recommendation: Leader)

Figure 8: Unifi Data Platform radar diagrams

Source: Ovum

Ovum SWOT Assessment

Strengths

A holistic platform for governance, cataloging, data prep, and workflows

The Unifi philosophy is that data prep needs to be an embedded part of the data governance and data integration process; to separate it would silo it, increase the potential for exposed sensitive data, and decrease the discoverability of data assets. Unifi's mission is to enable self-service; the Unifi Data

Platform supports four pillars of integrated functionality that span the data integration process leading up the analytics phase of data handling. These pillars are governance and security, catalog and discovery, data preparation, and workflow and scheduling. Bound by the platform's underlying engine of NLP, AI, and machine learning technology (OneMind), the whole is greater than the sum of its parts; all functionality is available in the same user-friendly ecosystem, enforced by the consistent data policies. Governance of data is especially strong, as controls for data are applied consistently throughout the platform and are not restricted just to the data prep process. The security frameworks and role-based access controls set in underlying data repositories are inherited and respected in the Unifi Data Platform, creating governance continuity across the enterprise IT ecosystem. Data cataloging capabilities, in particular, are strongly integrated and natively developed as part of the platform.

Integrates with an extremely broad array of BI and analytics tools

The Unifi Data Platform, by the company's own definition, is not meant to be a BI and analytics environment; the company is focused on the governance, integration, prep, and workflow aspects that lead up to the final analysis process. The company instead focuses on offering an extensive array of integrations with both leading and emerging analytics environments so that the enterprise can mix and match products depending on its specific needs. Integrations and/or partnerships are available for Tableau, Qlik, MicroStrategy, IBM, Microsoft, SAS, SAP, Looker, Arcadia Data, as well as other additional third-party BI and analytics tools. Visualizations from these analytics or data science tools can be embedded directly in the Unifi Data Platform to provide previews of visualization prior to committing to a large-scale data extract, further strengthening the integration and allowing for a seamless data prep and visualization experience for users.

Weaknesses

Visualization capabilities are powered exclusively via integrations

Unifi Data Platform has very few weaknesses in terms of core (or advanced) data prep functionality. The product provides all expected functionality for common manipulations of data required to join data sets, add derived column value, create data filters, and define aggregations. All of these are assisted by advanced AI-driven technology and recommendations, which makes user navigation simple. However, one area in which Unifi Data Platform differs from some other data prep tools is the lack of presence of native, interactive visualizations in which elements can be selected to dynamically prompt actions such as transformations. To be certain, this is a strategic decision that the company has made; it does not wish to overlap or compete in the visualization market. In fact, it is working with visualization partners such as Tableau to get even tighter integration. The Unifi Data Catalog features will be seamlessly embedded in Tableau in the next release of the product. Nevertheless, native visualizations for the exploration and discovery phase of the data prep process can potentially provide tighter integration with a product's data prep functionality, allowing users to prompt data prep functions, with a single click, directly from charts or graphs as they explore and profile data.

Lack of regional offices outside the US limits international growth

Unifi is a relatively small company compared to some of the other vendors that offer data prep capabilities, and the broad, multifunctional nature of its platform means that there are many moving parts in a typical deployment. Support is a critical component of implementation, not just for technical setup, but also for employee onboarding and training as well. However, as the number of Unifi customers grows, this support strategy will need to expand with its sights set on global horizons; the company has plans to expand internationally. As most existing customers are in North America and

Europe, operating out regional offices has, to date, been an effective strategy; the company's corporate headquarters is in San Mateo, California, and it also operates a development center in Bangalore, India. But with a growing number of customers located outside of the US, the company needs to expand with multiple international offices in order to continue to guarantee the level of support that it has become known for.

Opportunities

OneMind integrated AI capabilities will help expand the user base

In December 2017, Unifi announced OneMind: integrated AI capabilities that underpin the entire Unifi Data Platform. To be certain, the Unifi Data Platform previously had very strong AI and machine learning-driven functionality; what OneMind does is unify these capabilities into a single engine that drives recommendations and one-click capabilities for all four pillars of the platform's functionality. It also delivers advanced natural language query capabilities, allowing users to ask natural language questions directly of data and get immediate results. Capabilities for GDPR compliance, recommendations of workflow automation, and cost-based optimization of processing are all powered by OneMind. The net effect of the OneMind integrated AI functionality is that more users than ever can now drive business value from the Unifi Data Platform, regardless of their technical skill.

Strong governance capabilities position the platform for compliance

Unifi's holistic approach helps fulfill the needs for compliance by providing a single integrated environment where all self-service activities leading up to visualization and analysis can be controlled under the same framework. Many of Unifi's current customers came to the company initially with a compliance – rather than data prep – use case as their primary driver. Strong controls for governance, such as the ability to automatically help detect and classify sensitive data (such as SSN), encryption, data masking, complete audit trail, and mapping of data assets via cataloging all help the enterprise gain visibility and control over informational assets. These controls do not exist in isolation; the Unifi Data Platform also respects the role-based access controls and security frameworks set in underlying data repositories, such as Hadoop, ensuring for consistent data policies across the IT ecosystem.

Threats

BI and analytics partners are working on their own data prep initiatives

The market for self-service data prep capabilities and functionality is becoming increasingly "platformized." Many BI and analytics vendors, in particular, are rapidly adding in integrated data prep capabilities to their existing environments, making for a "stickier" environment for users where data prep can be done in the same interface as analysis. As these vendors build out their own data prep initiatives, it has the potential to threaten Unifi's strong network of analytics partnerships. In most cases, these analytics platform data prep capabilities are not as rich or as functional as those offered by established vendors such as Unifi or standalone data prep tools. For Unifi, the buffer against this trend will be demonstrating its value as a holistic platform that offers much more than just data prep.

Thriving market for third-party cataloging tools may threaten model

One of Unifi Data Platform's strengths is the tight, embedded integration of its natively developed data cataloging capabilities, which complement its data prep capabilities by giving users a comprehensive view of data sets, jobs, and workflows. However, it is impossible to ignore the growing market for standalone, neutral data cataloging tools that aim to provide the enterprise a centralized catalog for their entire IT ecosystem. This market of third-party tools could threaten Unifi's embedded model.

Appendix

Methodology

Vendors were invited to participate in this assessment based on the inclusion and exclusion criteria outlined at the outset of the report, and were included in the final assessment contingent on acceptance of invitation and completion of required responses. The report is structured around multiple-choice responses gathered from a detailed, structured questionnaire, which assesses more than 200 unique features related to self-service data prep functionality as well as measures of execution capabilities and revenue. Each vendor, as part of the research process, receives a comprehensive meeting and the opportunity to demo its product, if desired. Supplemental information is obtained from vendor literature and websites, as well as third-party sources and existing Ovum research. Before publishing, the report is peer-reviewed.

Further reading

Beyond Self-Serve: Expanding the End-User Audience of Data Prep, IT0014-003213 (January 2017)

SWOT Assessment: Alteryx Designer, 11.8, INT002-000084 (March 2018)

SWOT Assessment: ClearStory Data v2.7, INT002-000051 (December 2017)

SWOT Assessment: Datawatch Monarch, v14.3, INT002-000058 (January 2018)

SWOT Assessment: IBM Data Refinery, INT002-000085 (April 2018)

SWOT Assessment: Oracle Analytics Cloud for Data Prep, 17.4.5, INT002-000074 (February 2018)

SWOT Assessment: Trifacta Wrangler Enterprise, v4.2, INT002-000063 (January 2018)

SWOT Assessment: Unifi Data Platform, v2.6, INT002-000057 (January 2018)

On the Radar: Unifi Software provides a Swiss army knife for data lake integration and governance, IT0014-003331 (September 2017)

Author

Paige Bartley, Senior Analyst, Data and Enterprise Intelligence

paige.bartley@ovum.com

Ovum Consulting

We hope that this analysis will help you make informed and imaginative business decisions. If you have further requirements, Ovum's consulting team may be able to help you. For more information about Ovum's consulting capabilities, please contact us directly at consulting@ovum.com.

Copyright notice and disclaimer

CONTACT US

www.ovum.com
askananalyst@ovum.com

INTERNATIONAL OFFICES

Beijing
Dubai
Hong Kong
Hyderabad
Johannesburg
London
Melbourne
New York
San Francisco
Sao Paulo
Tokyo